P: ISSN NO.: 2321-290X RNI : UPBIL/2013/55327 VOL-5* ISSUE-1* September- 2017
E: ISSN NO.: 2349-980X **Shrinkhla Ek Shodhparak Vaicharik Patrika**

Numerical Simulation of Queue Length, Waiting Time, Load Distribution in Cloud System for Different Traffic and Job Scheduling Models

Abstract

The cloud computing represents a concept of networking where the computing resources and users are located at different places, and the resources (such as hardware, softwareetc.)are accessible to the cloud users as a service through a computer network. To maintain efficient utilization of its resources while providing better QoS (which includes minimization of pending jobs in queue, faster job execution); the load balancer uses different algorithms for job scheduling and resource utilization. In this paper we analyzed the different job scheduling algorithms used by load balancer in terms of Queue Length, Waiting Time, and Load Distribution for Different Traffic Models using Matlab.

R.K. Shrivastava
Professor,
Deptt.of Mathematics,
S.M.S Govt. Science Collage,
Jiwaji University,
Gwalior, M.P.

Vikash Goswami
Research Scholar,
Deptt.of Mathematics,
S.M.S Govt. Science Collage,
Jiwaji University,
Gwalior, M.P.

Keywords: Cloud Computing, Queue Management.

Introduction

In the terminology of computer networks, the cloud is defined as a cluster of computing resources, such as software, storage, processor, memory etc. The cloud serves these resources as a service to the users (consumers) through a network (most commonly using internet). The hiring of cloud resources reduces the financial requirements necessary for the development and maintenance of infrastructure.

The cloud is provided as a service, hence it must serve the users request quickly while efficiently utilizing the resources to keep operational cost as low as possible. The load balancer is responsible for effectively performing above. In simple terms the job of the load balancer is to distribute the traffic coming from different users on the available resources in such an optimal way that maintains the QoS while minimizes the operational cost.

While the job of load balancer is quite complex in this paper in we modeled it using queuing theory to evaluate the response time for services. The model can also be used to optimally scale the cloud system to guarantee the QoS for given response time, and efficient formation of VMs (virtual machines) based on the system load. We further evaluated the different scheduling models using numerical analysis and compare them with the queuing model derived.

The rest of the paper is organized as follows. The second section presents an overview on the related topic. The third section describes the cloud computing while the modeling is presented in fourth section. The simulation results are presented in fifth section followed by the conclusion in the sixth section.

Aim of the Study

This paper is aimed to (1) analyze the potential of applicability of queuing theory in the cloud system for the purpose of job scheduling and efficient resource utilization. (2) Development of accurate and reliable numerical simulation model for the analysis of complex queuing model development and analysis.

Review of Literature

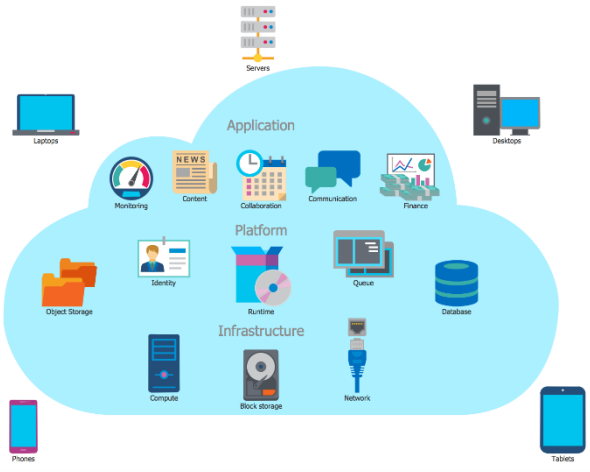
Vacation queuing theory based energy saving task scheduling in a heterogeneous cloud computing system is described by Chunling Cheng et al. [1] which deals in minimizing the power loss on the processing nodes remains powered on all the time to await incoming tasks, because the randomness in incoming jobs in cloud computing environments. Hamzeh Khazaei et al. [2] modeled a cloud system as an M/G/m queuing system, which means considering the inter-arrival time of requests is exponentially

distributed, and the service times of requests are i.i.d. random variables. Kaiqi Xiong et al. [3], modeled a cloud center as the classic open network; they assumed the inter-arrival time and obtained the distribution of response time. Using the distribution of response time, they also discovered the relationship among maximum number of tasks, highest level of services and the minimal service resources. Bo Yang et al.[4],computed the distribution of response time of the cloud center modeled as an M/M/m/N queuing system with finite buffer of size N and both Inter-arrival and service times are assumed to be exponentially distributed. Jordi Vilaplana et al.[5] estimated the QoS of cloud system modeled with an open Jackson network. The authors also determine and measure the guaranteed QoS the cloud can offer for the given response time. The analysis performed considering a number of different parameters, such as the job arrival rate and the number and service capacities of processing servers, etc .Gong et al. [6] analyzed the parallel processing in a non-dedicated distributed environment using a performance model to study the feasibility and limitations of parallel processing. It assumed the job process as an M/G/1 queuing system, wherein the parameters are estimated using simulation experiments. This model presented the detailed analysis of the task completion time. Zikos et al.[7] tested the job scheduling algorithm with heterogeneous servers and unknown service time. Two types of processors are considered in the cluster, with different performance and energy properties. The authors also proposed three local resource allocation policies, the highest performance policy, the probability policy, and the best energy deficiency policy, based on the M/M/n queuing model.

Cloud Computing

The term cloud computing is used to represent a cluster of computing resources, such as software, storage, processor, memory etc. which are provided as a service through a network. The name cloud comes from the cloud shaped symbol representing the complex infrastructure it contains.

Figure 1: Cloud computing logical diagram



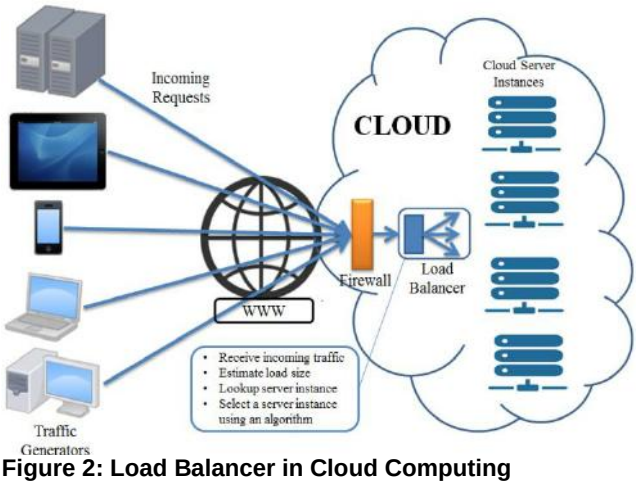


Figure 2: Load Balancer in Cloud Computing

Services on Cloud

1. Software as a Service (SaaS):this type of service offers both resources and applications. SaaS facilitates the user to use their software directly from cloud hence no need to install on your device [8].
2. Platform as a Service: this type of service offers access to the components that are required to develop and operate an application over the internet[8].
3. Infrastructure as a Service: this type of service offers the computational infrastructure including hardware and software [8].

Standard Load Balancing Algorithms

Round Robin Load Balancer

This is one of the simplest methods for distributing client requests across a group of VMs. In this algorithm the load balancer assigns the first client request to a VM picked randomly from the group and while further requests are assigned to other VMS in circular order.

Throttled Load Balancing Algorithm

This algorithm is based the selection of virtual machine best suitable for the clients requested task [9].

Equally Spread Current Execution Algorithm

This algorithm distribute the client requests to the lightest loaded virtual machine which can execute the assigned request easily and quickly to maximize throughput [10].

Cloud Modeling

Analytical Queuing Model for Cloud System

For the modeling we consider the cloud system as multi-VM system with the queuing model. All the requests in the cloud arrives through the load balancer (LB). The LB assigns the client requests to one of the VMs (virtual machines), the individual VM is denoted by VM_i , $i = 1,2,3 \ldots, m$ where m is the total VMs in cloud.

The load balancer is modeled as M/M/1queue, with both arrival rate (λ)and service rate (S)modeled as exponential probability distribution function where $\lambda \ll S$.

The load balancer selects a particular VM to assign received task on the basis of averaged workload in each VMs.

A VM represents the physical computational resources of the cloud system which contains all the necessary software's and hardware's required to execute the client's requests. All the VM's are assumed identical with arrival rate γ and service rate μ and modeled as an M/M/m queuing system.

The response time of the load balancer modeled as M/M/1 queuing can be given as

$$T_{LB} = \frac{1/S}{1 - \lambda/S} \quad (1)$$

Considering $S \gg \lambda$ also $S \gg 1$ the equation (1) results$T_{LB} \cong 0$.

The response time of the VMs considering there are at maximum$m$VMsin the cloud. Since VMs are modeled as M/M/m queue the response time can be given as

$$T_{VM} = \frac{1}{\mu} + \frac{C(m,\rho)}{m\mu - \gamma} \quad (2)$$

Where $C(m,\rho)$ represents Erlang's C formula to calculate the joining probability of a new client in M/M/m queue and given as

$$C(m,\rho) = \frac{\left(\frac{(m\rho)^m}{m!}\right)\left(\frac{1}{1-\rho}\right)}{\sum_{k=0}^{m-1}\left(\frac{(m\rho)^k}{k!}\right) + \left(\frac{(m\rho)^m}{m!}\right)\left(\frac{1}{1-\rho}\right)} \quad (3)$$

Where $\rho = \gamma/\mu$.

Numerical Model Considerations for Cloud System

Following consideration are taken into account when the numerical models is developed.

1. It is assumed that the load balancer knows the configuration (like processing capacity, memory etc.) of each virtual machine (VM) in the cloud.
2. Firstly the load balancer can get the operational state of each VM with zero time delay.
3. The load balancer takes no time in selecting and assigning the tasks to VM's.
4. The load balancer selects the VM for the input tasks on the basis of selected algorithm.
5. Each VM has zero booting time hence start executing assigned task immediately.
6. The incoming tasks size is considered in MI (million instructions) units.
7. The VM's capacities are also considered in MIPS (million instructions per second) units.
8. All the tasks length are integer multiple of the VMs capacity.

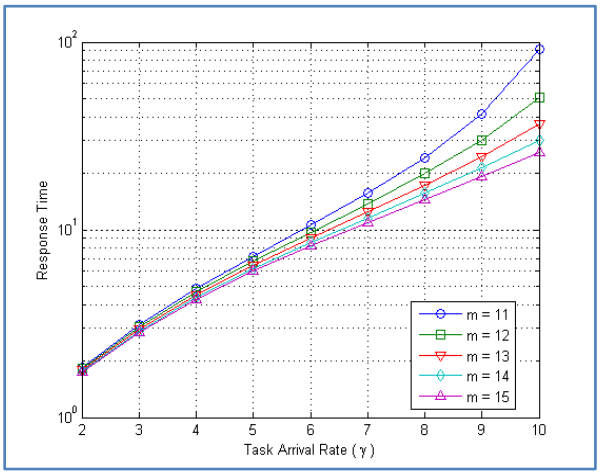
Simulation Results

The analytical and numerical simulations of all the systems are performed using MATLAB. The outcomes of the simulation for different configurations are presented usingplots and tables.

Analytical Model Results

The analytical simulation of the system is performed using the equation 2 and 3. This simulation shows the impact of increasing the number of VMs in the response time.

Figure 3: Response Time of the Cloud As A Function of Task Arrival Rate



The figure 3 shows how the response time is affected by increasing the task arrivalrate (λ) for different cloud size (number of VMs in cloud). The graph shows that initial increase in the number of VMsgreatly reduces the response time. However, the further increase in number of VMs does not reduce the response time at the same rate.

Numerical Simulation Results

The numerical modeling and simulation provides an efficient way to analyze the behavior of the systems that cannot be modeled analytically due to their complexity.

The numerical modeling of the system is performed for analyzing its behaviors for two different arrival rate distributions (exponential and normal) and two different balancing algorithms (round robin and equally spread).

For the numerical analysis the cloud is considered to have total 10 identical VMs with each having the capacity of 1 MIPS. The variation in task arrival rate in case of normal distribution is done by changing the standard deviation of the function.

To avoid abnormal deviation in results due to random nature of variables the same simulation is repeated for 50 time and then average of the values are taken as final results.

Figure 4: Response Time of the Cloud for Different Task Arrival Rate While the Task Arrival Rate is Following Exponential Probability Distribution

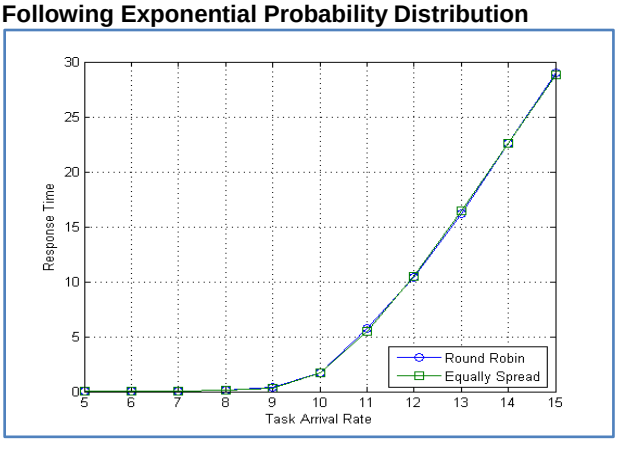


Figure 5: Individual VM Utilization Comparison for Different Load Balancing Algorithms While The Task Arrival Rate is Following Exponential Probability Distribution.

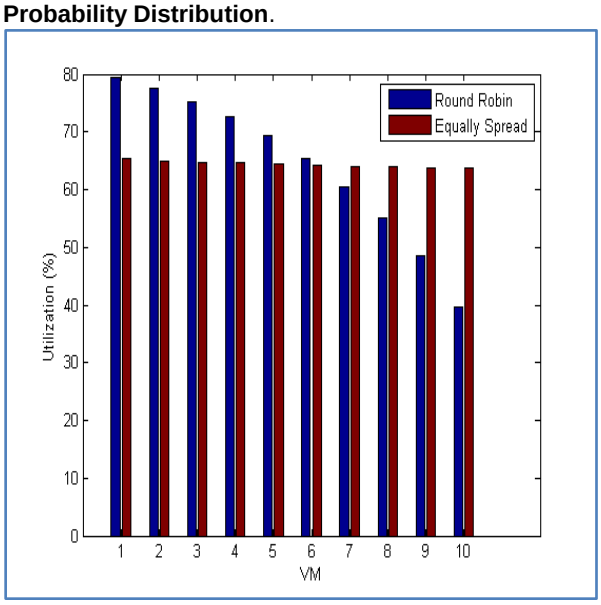


Figure 6: Response Time of The Cloud For Different Task Arrival Rate While The Task Arrival Rate Is Following Normal Probability Distribution And Arrival Rate Varied by Changing Standard Deviation of The Function.

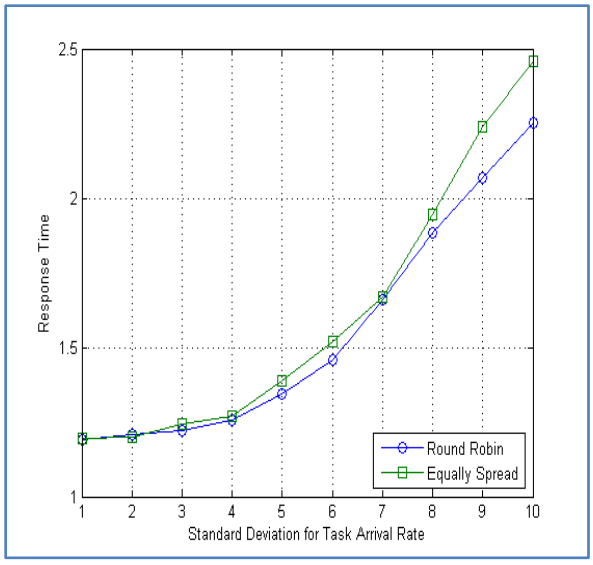
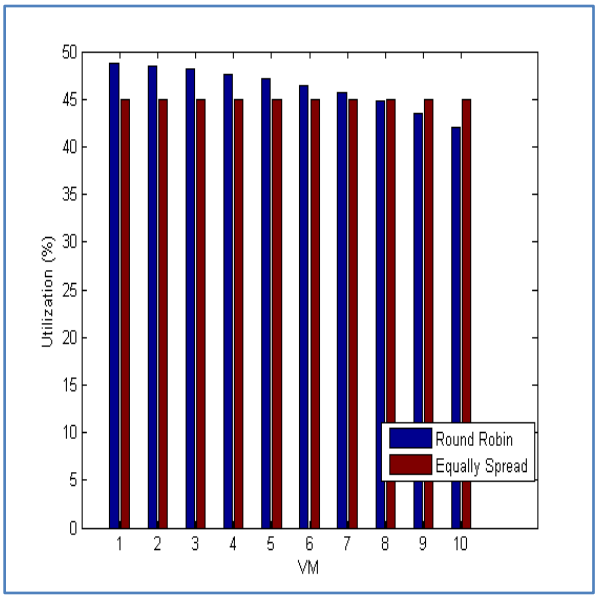


Figure 7: Individual VM Utilization Comparison for Different Load Balancing Algorithms While The Task Arrival Rate is Following Normal Probability Distribution And Arrival Rate Varied by Changing Standard Deviation of The Function.



Conclusion

This paper presents an analytical and numerical analysis of task scheduling and load balancing problem considering the queuing theory. From the simulation outcomes following conclusions can be drawn (1) for the task arrival rate which following the exponential probability distribution an exponential growth in response time can be seen with respect to task arrival rate, however the balancing algorithm has no impact on response time improvement. (2) For the similar task arrival the Equally Spread Current Execution Algorithm distributes the load uniformly over all available VMs, where in the Round Robin technique the maximally utilized VM shares the load approximately twice to that of minimally utilize VM. (3) For the task arrival rate which following the normal probability distribution the Round Robin technique provides better response time in comparison of Equally Spread Current Execution Algorithm. It can also be seen that for Round Robin response time even gets better for higher loads. (4) For the similar task arrival the tasks are distributed much uniformly over all available VMs even Round Robin the maximally utilized VM shares the load approximately 10% higher to that of minimally utilize VM.

References

1. *Chunling Cheng, Jun Li, and Ying Wang "An Energy-Saving Task Scheduling Strategy Based on VacationQueuing Theory in Cloud Computing", Tsinghua Science and TechnologyISSNll1007-0214ll04/11llpp28-39Volume 20, Number 1, February 2015*
2. *Hamzeh Khazaei, Jelena Misic and Vojislav B. Misic "Performance analysis of Cloud ComputingCenters", IEEE Transactions on*

Parallel and Distributed Systems, Volume: 23, Issue: 5, May 2012.

3. *Kaiqi Xiong and Harry Perros "Service Performance and Analysis in Cloud Computing", Services - I, 2009 World Conference on6-10 July 2009.*
4. *Bo Yang, Feng Tan, Yuan-Shun Dai "Performance evaluation of cloud service considering fault recovery", The Journal of Supercomputing July 2013, Volume 65, Issue 1, pp 426–444.*
5. *Jordi Vilaplana, Francesc Solsona, IvanTeixidó, Jordi Mateo, Francesc Abella, Josep Rius "A queuingtheorymodelforcloudcomputing", Science Direct: The Journal of SupercomputingJuly 2014, Volume 69, Issue 1, pp 492–507.*
6. *Linguo Gong, Xian-He Sun, Edward F. Watson"Performance Modeling and Prediction ofNondedicated Network Computing", IEEE Transactions on Computers, VOL. 51, NO. 9, September 2002.*
7. *Stylianos Zikos,Helen D. Karatza "Performance and energy aware cluster-level scheduling ofcompute-intensive jobs with unknown service times", Simulation Modelling Practice and Theory 19 (2011) 239–250.*
8. *Alexa Huth and James Cebula "The Basics of Cloud Computing", https://www.us-cert.gov/sites/default/files/.../CloudComputingHuthCebula.pdf.*
9. *Ms.NITIKA, Ms.SHAVETA, Mr. GAURAV RAJ "Comparative Analysis of Load Balancing Algorithms in Cloud Computing", International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May 2012.*
10. *MR.Manan D. Shah, MR.Amit A. Kariyani, MR.Dipak L. Agrawal "Allocation of Virtual Machines In Cloud ComputingUsing Load Balancing Algorithm", IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 3, No.1, February 2013.*